

Reply to Garcia et al.: Common mistakes in measuring frequency dependent word characteristics

Peter Sheridan Dodds,^{1,2,*} Eric M. Clark,^{1,2} Suma Desu,³ Morgan R. Frank,³ Andrew J. Reagan,^{1,2} Jake Ryland Williams,^{1,2} Lewis Mitchell,⁴ Kameron Decker Harris,⁵ Isabel M. Kloumann,⁶ James P. Bagrow,^{1,2} Karine Megerdooian,⁷ Matthew T. McMahon,⁷ Brian F. Tivnan,^{7,2,†} and Christopher M. Danforth^{1,2,‡}

¹*Computational Story Lab, Vermont Advanced Computing Core,*

& the Department of Mathematics and Statistics, University of Vermont, Burlington, VT, 05401

²*Vermont Complex Systems Center, University of Vermont, Burlington, VT, 05401*

³*Center for Computational Engineering, Massachusetts Institute of Technology, Cambridge, MA, 02139*

⁴*School of Mathematical Sciences, North Terrace Campus, The University of Adelaide, SA 5005, Australia*

⁵*Applied Mathematics, University of Washington,
Lewis Hall #202, Box 353925, Seattle, WA, 98195.*

⁶*Center for Applied Mathematics, Cornell University, Ithaca, NY, 14853.*

⁷*The MITRE Corporation, 7525 Colshire Drive, McLean, VA, 22102*

(Dated: May 29, 2015)

We demonstrate that the concerns expressed by Garcia *et al.* are misplaced, due to (1) a misreading of our findings in [1]; (2) a widespread failure to examine and present words in support of asserted summary quantities based on word usage frequencies; and (3) a range of misconceptions about word usage frequency, word rank, and expert-constructed word lists. In particular, we show that the English component of our study compares well statistically with two related surveys, that no survey design influence is apparent, and that estimates of measurement error do not explain the positivity biases reported in our work and that of others. We further demonstrate that for the frequency dependence of positivity—of which we explored the nuances in great detail in [1]—Garcia *et al.* did not perform a reanalysis of our data—they instead carried out an analysis of a different, statistically improper data set and introduced a nonlinearity before performing linear regression.

Note: The present manuscript is an elaboration of our short reply letter [2].

I. FUNCTION WORDS IN THE LIWC DATA SET ARE NOT EMOTIONALLY NEUTRAL

We first address Garcia *et al.*'s concerns about our online survey [3], which they suggest induced a positivity bias in respondents' answers.

Garcia *et al.* claim that a set of function words in the LIWC (Language Inquiry and Word Count) data set [4] show a wide spectrum of average happiness with positive skew (their Fig 1A) when, according to their interpretation, these words should exhibit a Dirac delta function located at neutral ($h_{\text{avg}}=5$ on a 1 to 9 scale). We expose and address two fundamental errors.

First, function words in the LIWC data set are simply not emotionally neutral. The LIWC data set annotates 4487 words and stems on a wide range of dimensions [4]. We find a total of 421 words and 48 stems are coded as function words with 450 matches in our data set when using stems. Of these, only 7 are indicated as emotional (5 positive, 2 negative) which appears to support Garcia *et al.*'s interpretation. However, a straightforward reading of the LIWC list of function words reveals that

these words readily bear emotional weight as exemplified by “greatest” and “worst”. We present some of the most extremely and most neutrally rated LIWC function words in Tab. I.

More generally, “Not looking at the words” and “Not showing the words” are pervasive issues with word- and phrase-based summary statistics for texts. We should be able to see how specific words contribute to summary statistics for texts to provide (1) an assurance the measure is performing as intended, and (2) insight into the text itself. For example, all sentiment scoring algorithms based on words and phrases must be able to plainly show why one text is more positive through changes in word frequency, such as through the word shifts we have developed for both print [5–7] and as interactive, online visualizations [8]. Elsewhere, in studying the Google Books corpus, we have produced analogous word shifts for the Jensen-Shannon divergence [9]. We exhort other researchers to produce similar word shifts (and not just word clouds), and to question work with no such counterpart.

Second, as we discuss in detail in Sec. III below, no statement about biases can be made about sets of words chosen without frequency of usage incorporated. Any given set of words may have a positive, neutral, or negative bias, but we must know how they are chosen before being able to generalize (as we have done thoroughly in [1]). Because we have no guarantee that the expert-generated LIWC function words are exhaustive and because we are merging words of highly variable usage frequency, a finding of an average positive bias for

* peter.dodds@uvm.edu

† btivnan@mitre.org

‡ chris.danforth@uvm.edu

| High h_{avg} | Neutral h_{avg} | Low h_{avg} |
|----------------|-------------------|----------------|
| billion 7.56 | been 5.04 | wouldnt 3.86 |
| million 7.38 | other 5.04 | not 3.86 |
| couple 7.30 | into 5.04 | shouldn't 3.84 |
| millions 7.26 | theyre 5.04 | none 3.84 |
| greatest 7.26 | it 5.02 | haven't 3.82 |
| rest 7.18 | some 5.02 | wouldn't 3.78 |
| best 7.18 | where 5.02 | fewer 3.72 |
| equality 7.08 | themselves 5.02 | lacking 3.71 |
| unique 6.98 | im 5.02 | won't 3.70 |
| plenty 6.98 | quarterly 5.02 | wasnt 3.70 |
| truly 6.86 | ive 5.02 | dont 3.70 |
| hopefully 6.84 | because 5.00 | don't 3.70 |
| first 6.82 | whereas 5.00 | down 3.66 |
| plus 6.76 | id 5.00 | nobody 3.64 |
| well 6.68 | til 5.00 | doesn't 3.62 |
| greater 6.68 | the 4.98 | couldnt 3.58 |
| highly 6.60 | to 4.98 | without 3.54 |
| me 6.58 | by 4.98 | no 3.48 |
| done 6.54 | or 4.98 | cant 3.48 |
| extra 6.52 | part 4.98 | zero 3.44 |
| infinite 6.44 | rather 4.98 | against 3.40 |
| simply 6.42 | its 4.96 | never 3.34 |
| equally 6.40 | when 4.96 | cannot 3.32 |
| sixteen 6.39 | perhaps 4.96 | lack 3.16 |
| we 6.38 | yall 4.96 | negative 2.42 |
| soon 6.34 | of 4.94 | worst 2.10 |

TABLE I. Three subsets of 450 LIWC function words with high, neutral, and low average happiness scores from our labMT study [1, 6] (stems provide more matches than those found by Garcia et al.). Each word’s score is the average rating for 50 participants (scale is 1 to 9 with 1 = most negative, 5 = neutral, and 9 = most positive). Function words may carry emotional weight and cannot be presumed to be neutral.

LIWC function words is meaningless, regardless of their transparent capacity for being non-neutral.

Emotional words in LIWC provide another case in point. Around 20% of the LIWC data set (907 words and stems) are denoted as having positive affect (160 words and 247 stems) or negative affect (151 words and 349 stems). While stems complicate word counts, the LIWC data set clearly does not show evidence of a positivity bias. Because the LIWC data set is expert-curated and meant to be general, it does not fit any natural corpora with respect to usage frequency (i.e., LIWC words constitute an unsystematic sampling). Word lists meant to accurately reflect statistical properties of language must be built directly from the most frequently used words of well defined corpora—a point we will return to several times in this reply. An earlier expert-curated word list, the smaller ANEW data set [10], similarly fails in these respects, showing a fairly flat distribution across average

happiness [5].

LIWC, along with all word data sets, should not be considered an unimpeachable “gold standard”; language is far too complex to make such an assured statement. All word data sets, including our own, will have limitations.

II. COMPARISON TO WARRINER AND KUPERMAN’S DATA SET

We next contend with a comparison made by Garcia *et al.* between our work on English with a similar sized survey by Warriner and Kuperman (WK) [11, 12]. WK generated a merged list of 13,915 English words, the bulk of which (11,826) are a list of lemmas taken from movie subtitles. Immediately, we have a mismatch: our word list incorporated the 5000 most frequently used words (or tokens) in each of four disparate corpora (New York Times, Google Books, music lyrics, and Twitter) whereas WK’s list is mostly lemmas (e.g., “sing” but not “sung” or “sang”) taken from one coherent corpus. Further, each word was scored by 50 participants in our study, compared with 14–20 for the WK study.

In their Fig. 1B, Garcia *et al.* show histograms for the two word lists, which seem to indicate more negative words in the WK list and a higher median word happiness for our word list. But such a comparison is unsound: the words behind each histogram are not the same and word frequency is not being controlled for. The two histograms cannot be sensibly compared, and we can discard Garcia *et al.*’s finding that the median level of average word happiness h_{avg} for our full data set is 0.28 above the median level for the WK data set.

Nevertheless, Garcia *et al.* do appropriately compare the shared subset of words found in both data sets, finding a much smaller difference between median values of h_{avg} of 0.07. They then suggest that our use of cartoon faces to indicate the 1 to 9 scale of happiness responses induces a positive bias in respondents’ choices, referencing a study that found a non-smiling face to be slightly negative [13]. Their claim lacks foundation for several reasons.

First, WK employed a reverse 9 point scale, with 1 = happy and 9 = unhappy, flipping the scores after completing the surveys (also used in [10]). We have no objection to WK’s approach but evidently this further complicates any comparison between the two studies. Indeed, one might reasonably hypothesize that flipping the direction of the ratings could be the sole cause of the minor discrepancy between the words scored by both studies.

Second, we gave all participants clear written instructions that 5 was neutral. In wanting to generate results that could be compared with existing work, we followed the design of Bradley and Lang in their ANEW study [10], who used both cartoon figures in their self-assessment manikins and written (spoken for ANEW) instructions; we departed only in orienting positive to the right. (As have many others, Garcia *et al.* have used the

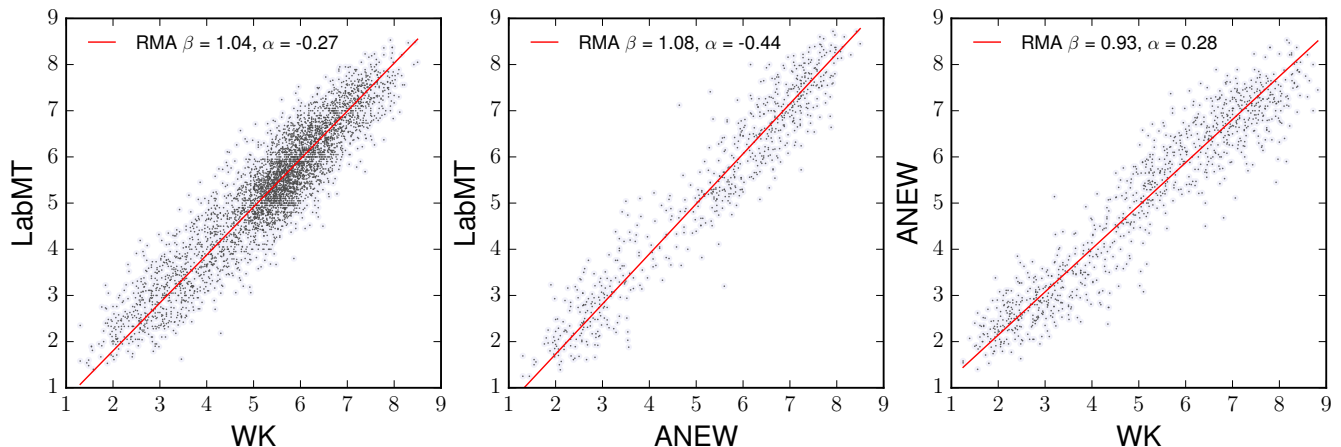


FIG. 1. Comparison of word ratings for three studies for overlapping words: labMT [1, 6], ANEW [10], and Warriner and Kuperman [11]. Reduced major axis regression [14] yield the fits $h'_{\text{avg}} = \beta h_{\text{avg}} + \alpha$.

ANEW study in their research [10, 15].) WK take pains to compare their scores with the ANEW study (which they use in part as a control) and other studies, finding their results are “roughly equivalent” (p. 6). And as we noted in [1], basic function words which are expected to be neutral such as “the” and “of” were appropriately scored as such, indicating that the survey mechanism we used was not adding a simple positive shift.

Third, given the nature of language and surveys and changing demographics online, an exact match for the medians would be a remarkable achievement. The agreement between the labMT (our English word set [1, 6, 16]) and WK is still a strong one, and we show scatter plots for the matching word happiness scores in Fig. 1 for labMT, ANEW, and WK. Visually, we see the three studies are sympathetic with each other, particularly when we acknowledge the typical standard deviation for the scores of individual words (on the order of 1 to 2). We used Reduced (or Standard) Major Axis regression [14] to obtain the fits shown in Fig. 1, $h'_{\text{avg}} = \beta h_{\text{avg}} + \alpha$. We see that ANEW’s scores grow slightly faster than that of both labMT and WK ($\beta = 1.08$ and 1.07) and WK similarly relates to labMT ($\beta = 1.04$).

Fourth, and rather finally, according to the argument of Garcia *et al.* regarding faces, the median for ANEW should be higher than that of WK (noting again that they used the same happy to sad directionality), yet we see the *opposite* (5.29 versus 5.44). Moreover, comparing medians alone is insufficient—our regression analysis shows that, for the words they have in common, WK appears more emotionally biased than labMT with $\beta = 1.04$. We note that a much richer comparison could be carried out at the level of individual ratings, but this is far too detailed for the present response.

III. DEPENDENCE OF POSITIVITY ON FREQUENCY OF USAGE

We turn now to Garcia *et al.*’s central claim: that we claimed to find that a positivity bias is *independent* of word frequency across 10 languages. In fact, we instead variously stated that a positivity bias is “strongly” and “largely” independent of frequency, and we explored the minor departures from pure independence in detail for all 24 corpora across 10 languages (see [1] and the paper’s online appendices).

Garcia *et al.* write that our paper specifically conflicts with two previous works, their own [15] and that of Warriner and Kuperman [12]. We are able to dismiss [15] due to it being founded on a misapplication of an information theoretic formula by Piantatosi *et al.* [17], and which we demonstrate elsewhere [18].

Setting aside this misrepresentation, Garcia *et al.*’s issue with our work becomes to what degree frequency independence is followed, and they provide an alternative analysis of how positivity behaves with usage frequency. Whereas we performed the regression $h_{\text{avg}} = \alpha r + \beta$ where r is rank, they claim $h_{\text{avg}} = \alpha \log_{10} f + \beta$ is more appropriate. Once again, we stand by our own principled analysis for the following reasons.

1. Mismatch of scored word list and word list with frequency: In attempting to say anything about a given quality of words as it relates to usage frequency within a specific corpora, a complete census of words by frequency must be on hand. Garcia *et al.* have taken our merged word lists for each language and applied them to data sets for which they do not necessarily fit. Problematically, their word lists do not contain ranks, and consequently there are words missing in uncontrolled ways from the data they perform regression on. For the example of English, our 10,222 words will (likely) match the most common words in any sufficiently large English corpus. But the matching becomes more peculiar the

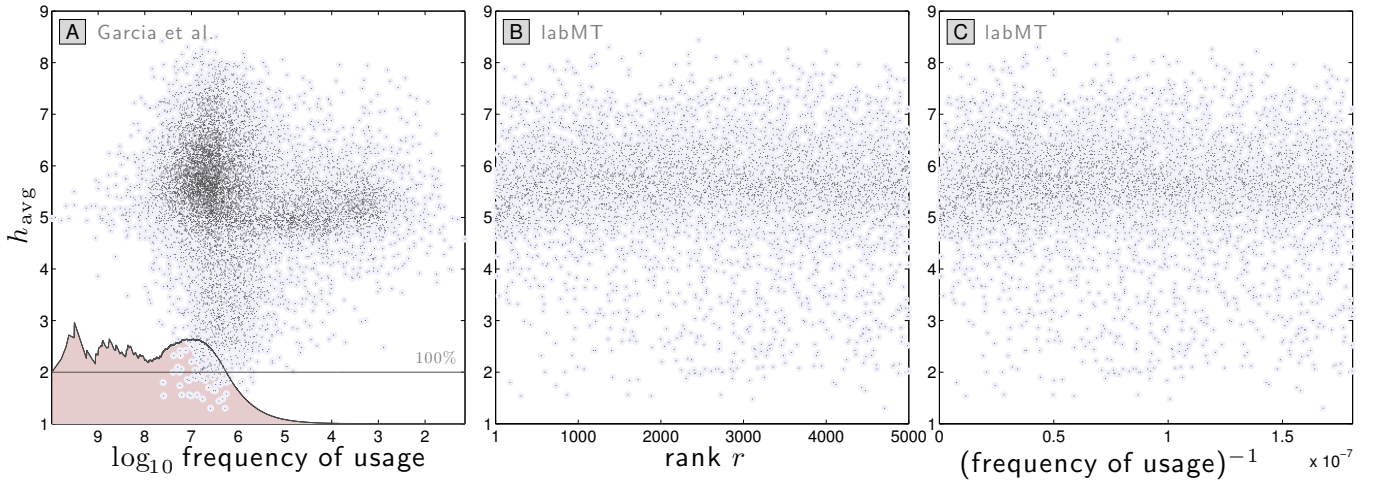


FIG. 2. **A.** Scatterplot of h_{avg} as a function of word usage frequency for the English Google Books word list generated by Garcia *et al.*. Uncontrolled subsampling of lower frequency words yields a lexicon that is not statistically representative of any natural language corpus. The lower curve provides a coarse estimate of cumulative lexicon coverage as a function of usage frequency f using Zipf’s law $f_r \sim f_1 r^{-1}$ inverted as $r \sim f_1/f_r$. The rapid drop off begins at around rank 5000, the involved lexicon size for Google Books in labMT [1, 6]. **B.** Scatterplot of h_{avg} as a function of rank r for the 5000 words for Google Books contributing to labMT, the basis of our jellyfish plots [1]. **C.** The same data as in **B** but now plotted against the inverse of usage frequency. The approximate adherence to Zipf’s law $f \sim r^{-1}$ means there is no substantive loss of information if regression is performed on the correct transformation of frequency. Linear regression fits for the first two scatterplots are $h_{\text{avg}} \simeq 0.089 \log_{10} f + 4.85$ and $h_{\text{avg}} \simeq -3.04 \times 10^{-5} r + 5.62$ (as reported in [1]). Note difference in signs, and the far weaker trend for the statistically appropriate regression against rank in **B**. Pearson correlation coefficients: +0.105, -0.042, and -0.043 with p -values 6.15×10^{-26} , 3.03×10^{-3} , and 2.57×10^{-3} . Spearman correlation coefficients: +0.201, -0.013, and -0.013 with p -values 6.37×10^{-92} , 0.350, and 0.350 (**B** and **C** must match). The Spearman analysis indicates that an assumption of a non-monotonic relationship between h_{avg} and rank r is well supported.

rarer the word, and the inclusion of Twitter in our word list means Garcia *et al.* have found “lolz”, “bieber” and “tweeps” in Google Books. In Fig. 2A, we plot average happiness as a function of frequency of usage for the word list they created from Google Books. The scatter plot is clearly unsuitable for linear regression. We show an estimate of cumulative coverage at the bottom (see caption), which crashes soon after reaching 5000 words.

2. Rank is an appropriate variable to regress happiness (or any word quality) against: Garcia *et al.* state that regression against frequency f is a better choice because information is lost in moving to rank r . However, the general adherence of natural language to Zipf’s law, $f \sim r^{-1}$, provides an immediate counterargument [19], even acknowledging the possibility of a scaling break [18]. Because word usage frequency is so variable, great care must be taken with any analysis. As we show for the case of English Google Books in Fig. 2A, regression on $\log_{10} f$ will be gravely compromised by the increasing preponderance of words at lower frequencies (a common issue with measuring power-law slopes), and, based on even the words for which coverage is reasonable, it would evidently be in poor judgment to extrapolate from any linear fit across frequencies. By

contrast, Fig. 2B shows how usage rank is perfectly suited for regression, and is the basis for the “jellyfish” plots we provided in Fig. 3 of [1] and in the paper’s online appendices. Our jellyfish plots make the general conformity to a rough (we do not claim “physical-law” strict) scale independence abundantly clear. By using rank, we are able to perform a much finer analysis than Garcia *et al.* propose, and we show in all corpora that the deciles for a sliding window of 375 ranks changes at most rather slowly. Finally, in Fig. 2C, we present how h_{avg} behaves as a function of $1/f$, illustrating both the error of choosing $\log_{10} f$ and that our results will be essentially unchanged if we regress against $1/f$.

In closing, we emphasize that minor deviations from frequency independence remain a secondary aspect of our observation that the Polyanna Hypothesis holds for a diverse set of languages, and are wholly irrelevant for the instrumental aspect of our work in creating text-based hedonometric tools.

ACKNOWLEDGMENTS

CMD was supported by NSF grant DMS-0940271; PSD was supported by NSF CAREER Award #0846668.

-
- [1] P. S. Dodds, E. M. Clark, S. Desu, M. R. Frank, A. J. Reagan, J. R. Williams, L. Mitchell, K. D. Harris, I. M. Kloumann, J. P. Bagrow, K. Megerdooimian, M. T. McMahon, B. F. Tivnan, and C. M. Danforth, *Proc. Natl. Acad. Sci.* **112**, 2389 (2015), available online at <http://www.pnas.org/content/112/8/2389>; online appendices: <http://compstorylab.org/share/papers/dodds2014a/>.
 - [2] P. S. Dodds, E. M. Clark, S. Desu, M. R. Frank, A. J. Reagan, J. R. Williams, L. Mitchell, K. D. Harris, I. M. Kloumann, J. P. Bagrow, K. Megerdooimian, M. T. McMahon, B. F. Tivnan, and C. M. Danforth, *Proc. Natl. Acad. Sci.* (2015), available online at <http://www.pnas.org/content/early/2015/05/20/1505647112>.
 - [3] D. Garcia, A. Garas, and F. Schweitzer, *Proc. Natl. Acad. Sci.* (2015), doi: 10.1073/pnas.1502909112.
 - [4] J. W. Pennebaker, R. J. Booth, and M. E. Francis, "Linguistic Inquiry and Word Count: LIWC 2007," at <http://bit.ly/S1Dk2L>, accessed May 15, 2014. (2007).
 - [5] P. S. Dodds and C. M. Danforth, *Journal of Happiness Studies* (2009), doi:10.1007/s10902-009-9150-9.
 - [6] P. S. Dodds, K. D. Harris, I. M. Kloumann, C. A. Bliss, and C. M. Danforth, *PLoS ONE* **6**, e26752 (2011).
 - [7] P. S. Dodds, E. M. Clark, S. Desu, M. R. Frank, A. J. Reagan, J. R. Williams, L. Mitchell, K. D. Harris, I. M. Kloumann, J. P. Bagrow, K. Megerdooimian, M. T. McMahon, B. F. Tivnan, and C. M. Danforth, *Proc. Natl. Acad. Sci.* **112**, 2389 (2015), available online at <http://www.pnas.org/content/112/8/2389>.
 - [8] hedonometer.org, accessed March 30, 2015.
 - [9] E. A. Pechenick, C. M. Danforth, and P. S. Dodds, "Characterizing the google books corpus: Strong limits to inferences of socio-cultural and linguistic evolution," (2015), available online at <http://arxiv.org/abs/1501.00960>.
 - [10] M. M. Bradley and P. J. Lang, *Affective norms for English words (ANEW): Stimuli, instruction manual and affective ratings*, Technical report C-1 (University of Florida, Gainesville, FL, 1999).
 - [11] A. B. Warriner, V. Kuperman, and M. Brysbaert, *Behav. Res. Methods* **45** (2013).
 - [12] A. B. Warriner and V. Kuperman, *Cognition & Emotion* (2014), published online October 14, 2014; <http://dx.doi.org/10.1080/02699931.2014.968098>.
 - [13] E. Lee, J. I. Kang, I. H. Park, J.-J. Kim, and S. K. An, *Psychiatry Research* **157**, 77 (2008).
 - [14] J. M. V. Rayner, *J. Zool. Lond. (A)* **206**, 415 (1985).
 - [15] D. Garcia, A. Garas, and F. Schweitzer, *EPJ Data Science* **1**, 3 (2012).
 - [16] I. M. Kloumann, C. M. Danforth, K. D. Harris, C. A. Bliss, and P. S. Dodds, *PLoS ONE* **7**, e29484 (2012).
 - [17] S. T. Piantadosi, H. Tily, and E. Gibson, *Proc. Natl. Acad. Sci.* **108**, 3526 (2011).
 - [18] J. R. Williams, J. P. Bagrow, C. M. Danforth, and P. S. Dodds, *Physical Review E* (2015), in press; Available online at <http://arxiv.org/abs/1409.3870>.
 - [19] G. K. Zipf, *Human Behaviour and the Principle of Least-Effort*, patterns (Addison-Wesley, Cambridge, MA, 1949).